Large-Scale Sequence and Structural Comparisons of Human Naïve and Antigen-Experienced Antibody Repertoires

Brandon J. DeKosky[1,#], Oana I. Lungu[1,4#], Daechan Park[1,4], Erik L. Johnson[1], Wissam Charab[1], Constantine Chrysostomou[1], Daisuke Kuroda[2], Andrew D. Ellington[3,4], Gregory C. Ippolito[4], Jeffrey J. Gray[2], George Georgiou[1,4,5,6,*]

SI Methods

Study Design

Sample collection size in repertoire subsets was dictated by biological sampling limitations and computational (time) throughput limits. A minimum of 500 computational models were generated per repertoire. All data were included and no outliers excluded unless otherwise stated in the particular statistical method.

Cell Isolation and VH:VL Sequencing

PBMC were isolated from donated human whole blood and non-B cells were depleted via magnetic bead sorting (Miltenyi Biotec, Auburn, CA). Approximately 70 mL whole blood was obtained for each individual. B cells were stained with anti-CD20-FITC (clone 2H7, BD Biosciences, Franklin Lakes, NJ, USA), anti-CD3-PerCP (HIT3a, BioLegend, San Diego, CA, USA), anti-CD19-v450 (HIB19, BD), anti-CD27-APC (M-T271, BD), and anti-IgD-PE (IA6-2, BD). $CD3^{-}CD19^{+}CD20^{+}CD27^{-}$ naïve B cells (NBCs) were analyzed for VH:VL sequences immediately following FACS sorting. $CD3^{-}CD19^{+}CD20^{+}CD27^{+}$ antigen-experienced B cells (AEBCs) (comprised of mostly memory B cells with a small number of peripheral plasmablasts) were incubated four days in the presence of RPMI-1640 supplemented with 10% FBS, 1× GlutaMAX, 1× non-essential amino acids, 1× sodium pyruvate and 1× penicillin/streptomycin (LifeTechnologies) along with 10 μg/mL anti-CD40 antibody (5C3, BioLegend), 1 μg/mL CpG ODN 2006 (Invivogen, San Diego, CA, USA), 100 units/mL IL-4, 100 units/mL IL-10, and 50 ng/mL IL-21 (PeproTech, Rocky Hill, NJ, USA) (1) prior to high-throughput VH:VL sequencing. High-throughput emulsion-based VH:VL sequencing was performed as reported previously (2). Briefly, cells were isolated into emulsion droplets along with poly(dT) magnetic beads for mRNA capture using a flow-focusing nozzle apparatus. Droplets contained lithium dodecyl sulfate and DTT to lyse cells and inactivate proteins, and mRNA released from lysed cells was captured by the poly(dT) sequences on

magnetic beads. The emulsion was broken chemically as described (2) and beads were collected, washed, and used as template for emulsion overlap extension RT-PCR which linked heavy and light chain transcripts into a single, linked cDNA construct for high-throughput sequencing via Illumina MiSeq 2x250 or 2x300 technology. Forward primers targeted the antibody Framework 1 regions (3); reverse primers targeted the IgM/Igκ/Igλ constant region for $CD27^-$ NBCs, and IgM/IgG/IgA/Igκ/Igλ reverse primers were used for $CD27^+$ AEBCs. Full length VH and VL genes were generated for antigen-experienced repertoires via bioinformatic assembly of three Illumina sequencing samples (VH:VL, VH only, and VL only) as described previously (2–4). The following barcoded primers were used for VH-only amplification and sequencing (barcodes are italicized): Donor 1 Replicate 1 5'-NNNN *TGAAGG* GGCTAGCTATTCCCATCGCGG-3', Donor 1 Replicate 2 5'-NNNN *CGCGTC* GGCTAGCTATTCCCATCGCGG-3', Donor 2 Replicate 1 5'-NNNN *TAAGAA* GGCTAGCTATTCCCATCGCGG-3', Donor 2 Replicate 2 5'-NNNN *AGCGAG* GGCTAGCTATTCCCATCGCGG-3'. The following barcoded primers were used for VL-only amplification and sequencing (barcodes are italicized): Donor 1 Replicate 1 5'-NNNN *TGAAGG* GCGCCGCGATGGGAAT-3', Donor 1 Replicate 2 5'-NNNN *CGCGTC* GCGCCGCGATGGGAAT-3', Donor 2 Replicate 1 5'-NNNN *TAAGAA* GCGCCGCGATGGGAAT-3', Donor 2 Replicate 2 5'-NNNN *AGCGAG* GCGCCGCGATGGGAAT-3'.

Bioinformatic Sequence Analysis

Illumina sequences were quality-filtered for a minimum Q-score of 20 over 50% of the raw reads. Reads were mapped to V-, D-, and J- genes and CDR3s extracted using both the International Immunogenetics Information System (IMGT) (5) and NCBI IgBlast software (6) with a CDR3 motif identification algorithm (7); the IMGT gene database version was current as of February 2014. Most antibody sequences were successfully mapped by both algorithms (96% of all sequenced antibodies), and IMGT gene assignments were given priority over IgBlast assignments; IMGT CDR3 length definitions

were used for genetic sequence analyses. Antibody gene isotype (G/A/M/K/L) was also assigned to each read based on the visible nested PCR primer sequences which target the constant region (2, 8). Sequence data were filtered for in-frame V(D)J junctions and productive VH and Vκ,λ sequences were paired by Illumina read ID and compiled by exact CDR3 junction nucleotide and V(D)J gene usage match. CDR-H3 junction nucleotide sequences were extracted and clustered to 96% nucleotide identity with terminal gaps ignored (USEARCH v5.2.32 (9)), with a minimum of one nucleotide mismatch permitted during CDR-H3 junction clustering regardless of sequence length, and the most abundant CDR-L3 corresponding to each CDR-H3 cluster seed was chosen as an H3:L3 pair. Resulting CDR-H3:CDR-L3 pairs with ≥2 reads comprised the preliminary list of VH:VL clusters for each data set. For determining germline identity in the FR3 region, all FR3 reads associated with the VH and VL in a given VH:VL pair were clustered by 90% identity using USEARCH, and the largest of the resulting clusters were analyzed by consensus sequence annotation using IMGT to determine percent homology to known germline genes. Naïve antibody sequences were filtered to include only those sequences with >98% germline identity in the FR3 region, similar to previous reports (10). Amino acid sequence hydrophobicity was determined using the normalized version of the Kyte-Doolittle hydrophobicity index (H-index) (11).

Raw DNA sequence data can be downloaded from the NCBI Short Read Archive (SRA) under accession numbers PRJNA315079, SRX709626 (Donor 1 antigen-experienced VH:VL), and SRX709625 (Donor 2 antigen-experienced VH:VL). Computer source code and associated data are available for download from the GitHub repository PNAS_2015-25510.

Structural Modeling

Antibody sequences represented by the most reads from Donor 1 and Donor 2 (with all selected antibodies being observed at >50 reads per sequence from the respective repertoire); naïve and antigen-experienced sets were analyzed. Antibody sequences were tested for uniqueness in and across repertoires, so that no antibody was modeled more than once. Antibodies with a CDR-H3 length of ≥16 (Chothia

numbering) amino acids were excluded from modeling. All sequences were subsequently filtered to ensure that each FR and CDR was identifiable by the modified Chothia definitions. Antibodies for which high sequence identity templates were available for CDR-H1, CDR-H2, CDR-L1 CDR-L2 and CDR-L3 were input through the RosettaAntibody 3.0 antibody modeling protocol as described (12). First, high-quality crystal structure templates (resolution $\leq$ 2.8 Å; CDR Cα B-factors $\leq$ 50) with high sequence identity for each FR (FR1-4) and CDR (1-3) were selected via BLAST searches and grafter together. The grafted models were then refined via backbone and sidechain minimization, repacking, and relaxation (13). The CDR-H3 loops of successfully refined models were then modeled *de novo* using the Next-Generation KIC algorithm while simultaneously refining V_H:V_L orientations via Rosetta SnugDock (14). Those CDR-H3 loops predicted to adopt a kinked conformation were *de novo* modeled with constraints, restricting the pseudodihedral angle of the four consecutive Cα atoms of residues H100X, H101, H102, and H103 to -10º to 70º. A total of 1,000 trajectories were modeled per antibody, with the lowest scoring models, as evaluated by the Rosetta scoring function, being chosen for visual inspection and further analysis.

Gene usage was highly correlated in sequence and structural repertoires, with Spearman ρ=0.84 for naïve repertoires in IGHV gene usage; ρ=0.91 for antigen-experienced repertoires in IGHV gene usage; ρ=0.88 for naïve repertoires in IGKV gene usage; ρ=0.91 for antigen-experienced repertoires in IGKV gene usage; ρ=0.87 for naïve repertoires in IGLV gene usage; and ρ=0.68 for antigen-experienced repertoires in IGLV gene usage.

Paratope Analysis

The CR-paratope comprised residues that were part of the contact region of each antibody as defined by Stave et *al*. (15). These consisted of V_H residues number 26-33 (CDR-H1); 50-58 (CDR-H2); 94-101 (CDR-H3); and V_L residues 27 to 32 (CDR-L1); 49 to 56 (CDR-L2); 91 to 96 (CDR-L3) in the Chothia numbering scheme. CR-paratope solvent accessible surface area (SASA) and hydrophobic

solvent accessible surface area (hSASA) were calculated using Rosetta (16). Charge was calculated from the number of negative (D and E) and positive (K, R) residues on the putative CR-paratope (17). The set of 141 non-redundant (by unique CDR-H3 sequence) human antibody crystal structures with a resolution of <4Å as of April 2014 which passed all structural modeling filters applied to repertoire sets were subjected to CR-paratope analysis side-by side with computationally modeled repertoires to ensure physically relevant metrics.

Antibody Framework Analysis

Similarities between FRs (FR1-3) of antibodies were calculated by determining the root-mean square deviation (RMSD) over the backbone atoms (C, Cα, N, O) of each antibody FR1-3 region to all other antibodies in a repertoire using the McLachlan algorithm (18) as implemented in the ProFit software (Martin ACR, Porter CT. Available at: http://www.bioinf.org.uk/software/profit/). Antibodies were then grouped by IGHV gene usage (same gene, same family, or different family); median RMSD values, standard deviations, and statistical significance of distributions were determined using R 3.1.1. V-gene segments which were utilized in <2 antibody models in both naïve and antigen-experienced repertoires in a particular donor were excluded from RSMD analysis for that particular donor. Framework residues used in the analysis were V_H: 8-25, 36-51, 57-94 and V_L: 10-23, 35-49, 57-88 in Chothia numbering.

Control Antibody Datasets

A set of 141 non-redundant human antibody crystal structures with a resolution of <4Å was selected for passing all structural modeling filters and subjected to CR-paratope analysis side-by side with computationally modeled repertoires to ensure physically relevant metrics. PDB codes for this dataset: 1ad0 1ad9 1adq 1aqk 1bbj 1bey 1bvk 1dee 1dl7 1dn0 1dql 1fvd 1gaf 1h0d 1i9r 1igm 1iqd 1jpt 1jv5 1kfa 1l7i 1mco 1mim 1nl0 1rz7 1t3f 1u6a 1uj3 1vge 1w72 1wt5 1za6 2agj 2aj3 2b1h 2cmr 2d7t 2eiz 2fee 2fjh 2fl5 2g75 2h9g 2j6e 2jb5 2qqk 2qqn 2qr0 2qsc 2r0k 2r56 2uzi 2vxq 2vxv 2wuc 2xa8 2xra 2xzc 2yc 2zkh

3aaz 3d85 3dgg 3dif 3dvg 3eo9 3g04 3g6a 3giz 3gjf 3gkw 3go1 3h0t 3h42 3hc3 3hc4 3hi6 3hmx 3k2u 3kdm 3kr3 3kym 3l5y 3ma9 3mlr 3mxw 3n85 3n9g 3na9 3ncj 3nfs 3nh7 3oaz 3p0y 3pgf 3qcu 3qot 3r1g 3sdy 3se9 3sqo 3t2n 3tnm 3tnn 3u0t 3u30 3uji 3ujj 3uls 3wd5 4d9l 4d9q 4dag 4dgy 4dke 4dkf 4dtg 4fqi 4g3y 4g5z 4g6a 4g6m 4gsd 4hcr 4hfu 4hfw 4hg4 4hie 4hj0 4hpy 4hs6 4hs8 4i77 4j6r 4jam 4jpi 4jzn 4ky1 4lmq 4lst 4lsu.

Naïve Antibody Modeling Control Dataset

RosettaAntibody models have previously been reported to have an accuracy of 1 Å RMSD in framework and canonical loops, and 2 Å in CDR-H3 loops. Independent analysis of seven germline antibodies (containing 100% sequence identity to germline gene segments in V-genes) from the PDB was performed. Antibodies in this set were computationally re-modeled, excluding homologs (antibody.py –exclude_homologs flag), and their RMSDs over backbone atoms were compared to those of the crystal structures using the McLachlan algorithm as implemented in the ProFit software. This analysis yielded average RMSDs of FRH: 0.744Å; CDR-H1:1.18 Å: CDR-H2: 1.40 Å; CDR-H3: 2.42 Å; FRL: 0.64 Å; CDR-L1: 0.83 Å; CDR-L2: 0.82 Å; CDR-L3: 1.02 Å. Antibodies included in this dataset were: 2XZA, 3EYQ, 3F12, 3QOS, 3QOT, 4JPI, 4JDV. CDR-H3s were not compared for 3EYQ, 3F12, due to this segment missing in PDB loops.

Statistical Analysis

Pearson Hierarchical Clustering: R (version 3.1.1) was used for hierarchical clustering (function "hclust"). The fractional frequency of V-gene pairs was first multiplied by a scaling factor of 100,000. After discarding gene pairs with zero fraction, fractions were log2-transformed and normal distributions were generated. Distance between samples was measured by Pearson correlation with complete-linkage as the agglomerative method.

Principal Component Analysis (PCA): PCA (the "princomp" function in MATLAB R2012b) was applied to processed Pearson hierarchical clustering data.

Linear Models for Microarray Data: R (version 3.1.1) was used for the identification of differentially paired genes (package "limma" version 3.14.4) (19, 20). Although the Linear Models for Microarray Data method (limma) was originally developed to identify differentially expressed genes in microarray data, the algorithm is applicable to quantitative PCR or RNA-Seq that provides a matrix composed of genes and expression values, and the linear model-based test is stable for experiments with a small number of replicates in that it borrows information across genes. Before running limma, gene pairs with zero usage were removed and quantile normalization was performed to normalize the difference in distribution of values among samples. *p*-values for multiple comparisons were corrected with the Benjamini-Hochberg procedure. Differentially paired gene cut-offs were established at a fold-change of 2 and an adjusted *p*-value of 0.05.

Kolmogorov-Smirnov (K-S) Test: The K-S test is a nonparametric test that compares the equality of two distributions. R (version 3.1.1) was used for K-S statistical analyses (function "ks.test"). Raw values such as charge, length, hydrophobicity were used to compare the probability distributions across experimental groups.

Z-score: Z-score was used to compare two proportions of amino acid charges. The two-tailed *p*-value was computed using the pnorm function in R.

RosettaAntibody 3.0 Modeling Flags

Grafting

./antibody.py --light-chain <L.fasta> --heavy-chain <H.fasta>

Constraints

Dihedral CA 214 CA 215 CA 216 CA 217 SQUARE_WELL2 2.704 0.523 100

CDR-H3 *de novo* loop modeling

Executable: ./antibody_H3.mklmpi.linuxiccrelease
Flags:
-nstruct 1000
-s grafting/model.pdb
-constraints:cst_file grafting/cter_constraint
-antibody::remodel perturb_kic
-antibody::snugfit true
-antibody::refine refine_kic
-antibody::cter_insert false
-antibody::flank_residue_min true
-antibody::bad_nter false
-antibody::h3_filter false
-antibody::h3_filter_tolerance 5
-antibody:constrain_cter
-antibody:constrain_vlvh_qq
-ex1
-ex2
-extrachi_cutoff 0
-corrections:score:use_bicubic_interpolation false
-restore_pre_talaris_2013_behavior
-loops:legacy_kic false
-loops:kic_min_after_repack true
-loops:kic_omega_sampling
-loops:allow_omega_move true
-kic_bump_overlap_factor 0.36
-loops:ramp_fa_rep
-loops:ramp_rama
-loops:outer_cycles 5

SI Methods References

1. Recher M, et al. (2011) IL-21 is the primary common γ chain-binding cytokine required for human B-cell differentiation in vivo. *Blood* 118(26):6824–6835.
2. DeKosky BJ, et al. (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* 21(1):86–91.
3. DeKosky BJ, et al. (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotech* 31(2):166–169.
4. Lavinder JJ, et al. (2014) Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A* 111(6):2259–2264.
5. Brochet X, Lefranc M-P, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 36(suppl 2):W503–W508.
6. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* 41(W1):W34–W40.
7. Ippolito GC, et al. (2012) Antibody repertoires in humanized NOD-scid-IL2R gamma(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS ONE* 7(4):e35497.
8. McDaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G (2016) Ultra- high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat Protoc* 11(3):429–42.
9. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.
10. Glanville J, et al. (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc Natl Acad Sci U S A* 108(50):20066–20071.
11. Eisenberg D (1984) Three-dimensional structure of membrane and surface proteins. *Annu Rev Biochem* 53(1):595–623.
12. Weitzner BD, Kuroda D, Marze N, Xu J, Gray JJ (2014) Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins* 82(8):1611–1623.
13. Bradley P (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309(5742):1868–1871.
14. Sircar A, Gray JJ (2010) SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol* 6(1):e1000644.
15. Stave JW, Lindpaintner K (2013) Antibody and antigen contact residues define epitope and paratope size and structure. *J Immunol* 191(3):1428–35.
16. Leaver-Fay A, et al. (2011) Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487:545–574.
17. Der BS, et al. (2013) Alternative computational protocols for supercharging protein surfaces for reversible unfolding and retention of stability. *PLoS ONE* 8(5):e64363.
18. McLachlan AD (1982) Rapid comparison of protein structures. *Acta Crystallogr A* 38(6):871–873.
19. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments: statistical applications in genetics and molecular biology. *Stat Appl Genet Mol Biol* 3(1):Article 3.
20. Smyth GK (2005) limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health., eds Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S (Springer New York), pp 397–420.

Supplementary Tables

Table S1. Number of unique sequences and structural models encompassing naïve and antigen-experienced B cell repertoires in each donor.

Donor	*$CD3^{-}CD19^{+}CD20^{+}$ $CD27^{-}$ Naïve*	*Naïve Models*	*$CD3^{-}CD19^{+}CD20^{+}CD27^{+}$ Antigen-Experienced*	*Antigen-Experienced Models*
1	13,780	505	34,692	502
2	26,372	509	89,249	513
3	15,203	-	-	-
Total	55,355	1,014	123,941	1,015

Table S2. Heavy:light V-gene pairs with differential expression frequencies in NBC vs. AEBC B-cell receptor repertoires. Statistically significant differentially expressed heavy/light V-gene pairs with adjusted $p < 0.05$ between NBC and AEBC repertoires are listed here. Positive fold-change values denote VH:VL gene pairs that were more frequent in antigen-experienced datasets. See also Figure 2C. (Abbreviations: log FC indicates $\log_2$ fold change between conditions; Average Frequency indicates $\log_2$ average frequency across all observed values; and Adjusted P Value indicates *p* after multiple test correction.)

Gene Pair	*Log FC*	*Average Frequency*	*Adjusted P Value*
HV3-33:KV1-8	-4.625	2.934	2.68E-03
HV6-1:KV1-33	-4.510	3.723	2.68E-03
HV4-34:KV1-8	-4.127	3.796	8.46E-03
HV3-74:KV4-1	3.986	3.933	8.46E-03
HV3-74:KV2-28	4.151	3.044	8.46E-03
HV6-1:LV3-19	-3.624	2.692	1.38E-02
HV3-74:LV2-8	3.510	2.410	1.38E-02
HV1-69:KV1-8	-3.952	3.334	1.38E-02
HV3-7:KV4-1	3.595	4.312	1.38E-02
HV3-15:KV2-28	3.455	3.317	1.38E-02
HV1-18:KV1-8	-3.454	3.035	1.38E-02
HV3-74:LV1-51	3.289	2.490	1.59E-02
HV4-59:KV1-8	-3.292	4.021	1.59E-02
HV6-1:LV1-40	-3.251	3.659	1.59E-02
HV1-24:LV2-14	-3.250	4.539	1.59E-02
HV1-58:KV1-33	-3.039	2.682	2.49E-02
HV4-34:LV3-1	-2.907	4.976	3.14E-02
HV5-51:KV1-8	-3.339	2.720	3.14E-02
HV1-3:KV4-1	3.324	3.625	3.14E-02
HV1-58:LV3-1	-2.873	2.283	3.14E-02
HV3-30:KV1-8	-2.917	3.613	3.14E-02
HV6-1:LV3-1	-3.043	3.618	3.14E-02
HV1-58:KV1-39	-3.768	3.078	3.14E-02
HV4-61:KV3-20	2.963	3.795	3.99E-02
HV1-69:LV3-1	-2.638	5.963	3.99E-02
HV3-21:KV1-8	-2.767	3.706	4.35E-02
HV2-26:KV1-33	-2.606	3.537	4.35E-02
HV1-46:KV1-33	-2.554	5.407	4.79E-02

Table S3. Gene usage, amino acid, and nucleotide sequence details for five public antigen-experienced exact amino acid match CDR-H3:CDR-L3 antibodies. Non-templated bases are in bold uppercase, and mutations from germline underlined; differences between pairs are highlighted in gray. Different nucleotide sequences (including disparate non-templated bases resulting from N/P addition) and heavy chain isotypes indicated that these public antibodies originated from distinct heavy and light chain recombination events. D – Donor

D	CDR-H3_aa	CDR-L3_aa	CDR-H3_nt	CDR-L3_nt	Genes & Isotype
1	CARTARLLDYW	CMQGTHWPFTF	tgtgcgaga**ACTGCGA**ggctacttgactactgg	tgcatgcaaggtacacactggccattcactttc	HV3-11:D5-12:J4 κV3-20:J3 IgA
2	CARTARLLDYW	CMQGTHWPFTF	tgtgcgaga**ACCGCACG**gctgct**G**gactactgg	tgcatgcaaggtacacactggcc**GTT**cactttt	HV4-59:D2-15:J4 κV3-20:J2 IgM
1	CAKGSNWGSGYYFDYW	CQQYNYYPITF	tgtgcgaaag**GCTCGAA**ttgggg**TTCCGGGT**act actttgactactgg	tgccagcagtataatt<u>a</u>ttacccgatcaccttc	HV3-23:D3-16:J4 κV1-16:J5 IgM
2	CAKGSNWGSGYYFDYW	CQQYNYYPITF	tgtgcgaaag**GTT**ctaactggggga**TCCGGAT**act actttgactattgg	tgccaacagtataatt<u>a</u>ttacccgatcaccttc	HV3-23:D7-27:J4 κV1-16:J5 IgM
1	CARTNGYLDYW	CAAWDGSLNGWVF	tgtgcaagaa**CAAA**tggttat**C**ttgactactgg	tgtgcagcatgggatg<u>g</u>cagcctgaatggttgggtgttc	HV6-1:D3-3:J4 λV1-44:J3 IgM
2	CARTNGYLDYW	CAAWDGSLNGWVF	tgtgcaagaactaa**CGGG**tacttggactactgg	tgtgcagcatgggatg<u>g</u>cagcctgaatggttgggtgttc	HV6-1:D2-8:J4 λV1-44:J3 IgM
1	CTRGGLGTGIDYW	CTQATQFPYTF	tgt<u>a</u>caagag**GGGGT**ctggggaCCggtattgact actgg	tgca<u>c</u>gcaagctacacaatttccgtacacttttt	HV3-74:D2-21:J4 κV2-24:J2 IgG
2	CTRGGLGTGIDYW	CTQATQFPYTF	tgt<u>a</u>c<u>a</u>ag<u>a</u>ggtggcctgggga**CCGGGA**ttgact attgg	tgca<u>c</u>gcaagctacacaatttccgtacacttttt	HV3-23:D3-10:J4 κV2-24:J2 IgM
1	CAGDYGSGSYRFDYW	CMQGTHWPLTF	tgtgcg**GGGGAT**tatggttc**AGGC**agttatcg**A**t ttgactactgg	tgcatgca<u>g</u>ggtacacactggcctctcacttc	HV4-30-2:D3-10&3-16:J4/κV2-30:J4 IgM
2	CAGDYGSGSYRFDYW	CMQGTHWPLTF	tgtgcg**GGGGAT**tatggttcggg<u>a</u>agttat**CG**ct ttgactactgg	tgcatgcaaggtacacactggcc**CC**tcaccttc	HV4-30-2:D3-10:J4 κV2-30:J5 IgM

Table S4. Charge distribution mean values. Repertoire charge means, subdivided by H3:L3 total charge, H3 charge, L3 charge, and CR-paratope charge.

	Naïve				Antigen-Experienced		
	Donor 1	*Donor 2*	*Donor 3*	***Average***	*Donor 1*	*Donor 2*	***Average***
H3:L3 Total Charge	-0.291	-0.619	-0.366	**-0.467**	-0.0912	-0.102	**-0.0992**
H3 Charge	-0.215	-0.357	-0.324	**-0.312**	-0.0767	-0.0906	**-0.0867**
L3 Charge	-0.0760	-0.262	-0.0422	**-0.155**	-0.0145	-0.0117	**-0.0125**
CR-Paratope Charge	-0.954	-1.27	-	**-1.11**	-0.667	-0.684	**-0.676**

Table S5. Mean values for Igκ and Igλ subsets. CDR-L3 charge and average hydrophobicity index (H-index) means derived from histogram data presented in Figure S9. Igκ CDR-L3s showed more positive charge and more negative average hydrophobicity than Igλ CDR3s (see Figure S9).

	Igκ			**Igλ**		
	Naïve	*Ag-Exp*	*Difference*	*Naïve*	*Ag-Exp*	*Difference*
CDR-L3 Charge	0.132	0.340	0.208	-0.471	-0.430	0.041
CDR-L3 Avg H-index	-0.303	-0.266	0.037	0.111	0.127	0.016

Supplementary Figures

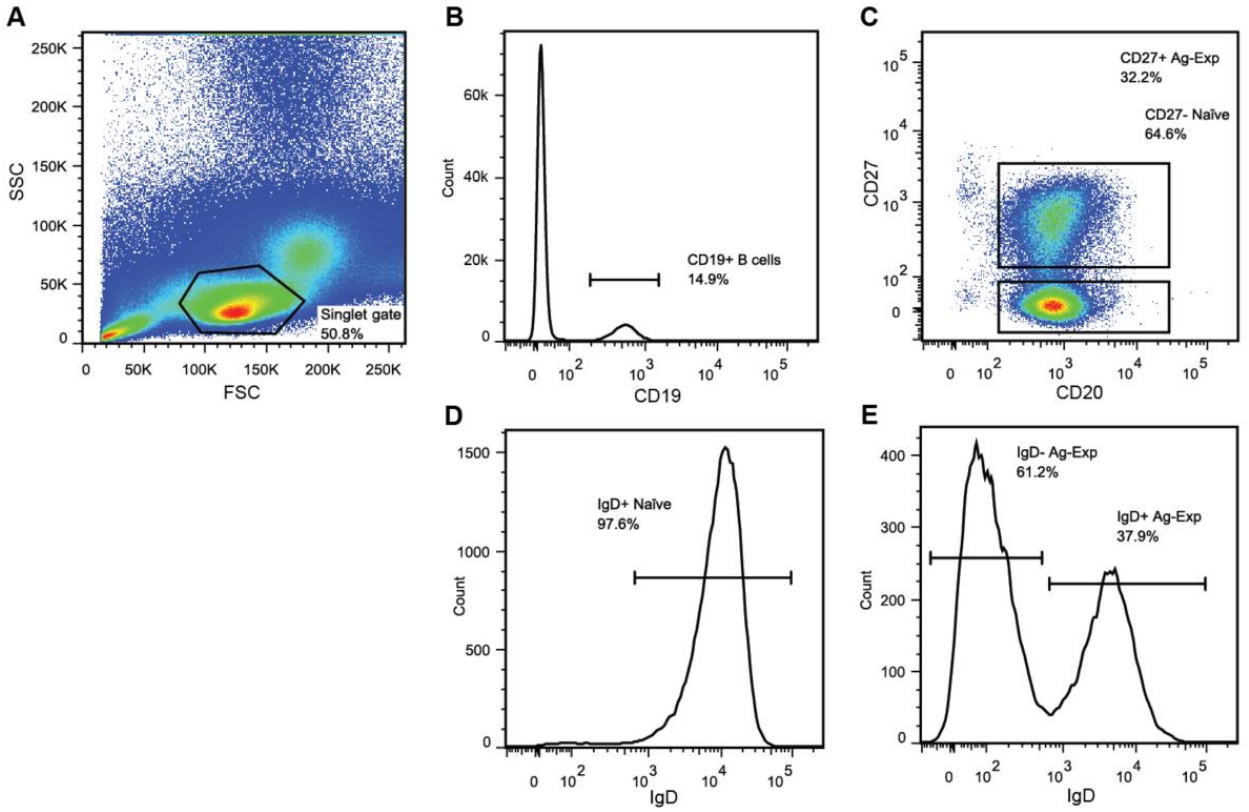

Fig. S1. FACS sorting of naïve and antigen-experienced B cell subsets from three human donors. *(A)* Singlet lymphocytes gate, *(B)* $CD19^+$ gate, *(C)* $CD20^+CD27^-$ and $CD20^+CD27^+$ gates. *(D)* IgD expression of $CD19^+CD20^+CD27^-$ NBCs. *(E)* IgD expression of $CD19^+CD20^+CD27^+$ AEBCs.

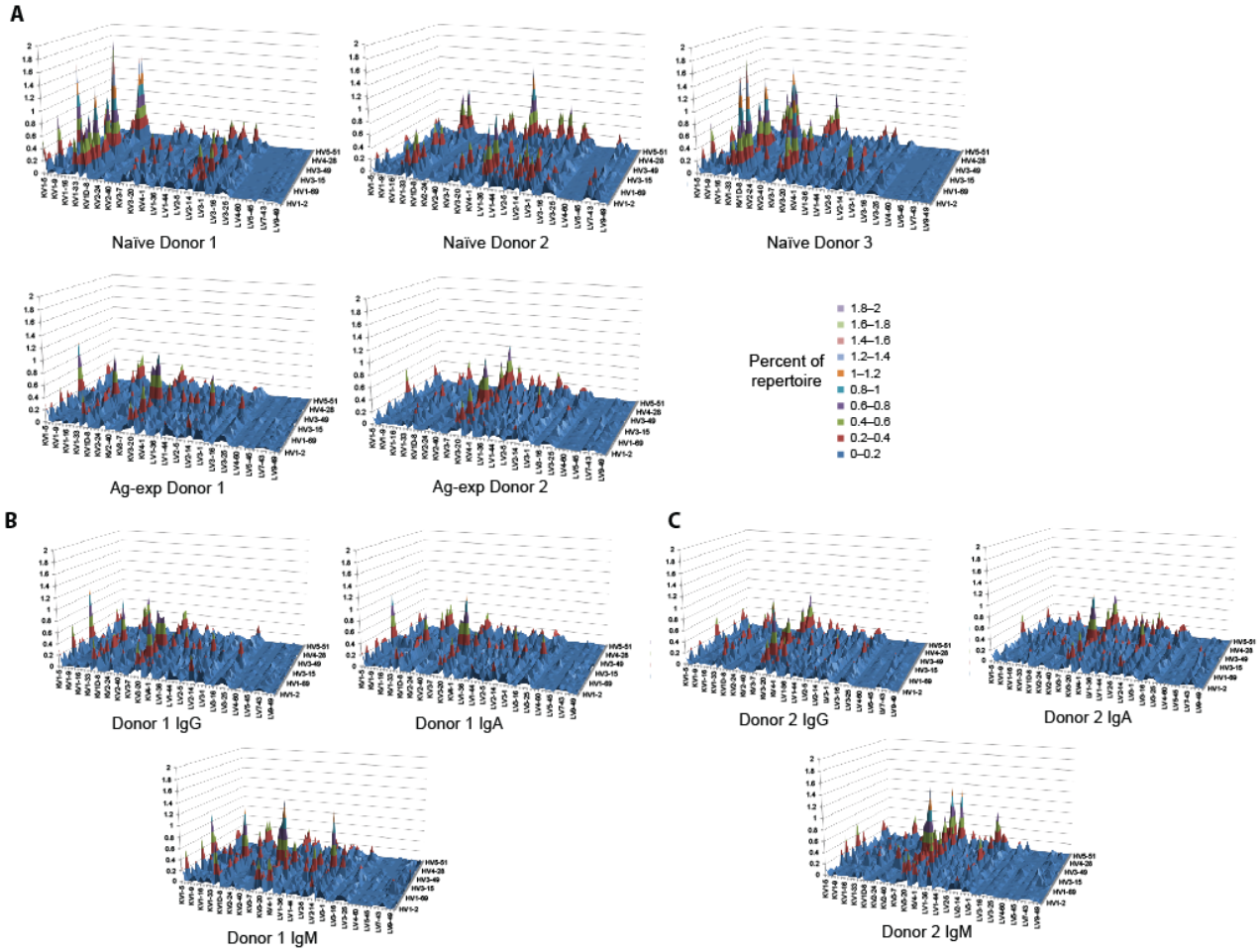

Fig. S2. Paired heavy:light V-gene usage surface maps of sequenced antibody repertoires. *(A)* V-gene surface maps shown for all sequenced naïve and antigen-experienced (Ag-exp) donor repertoires. *(B)* and *(C)* V-gene pairing surface maps for antigen-experienced B-cell receptors sequenced in Donor 1 *(B)* and Donor 2 *(C)*, separated by heavy chain isotype.

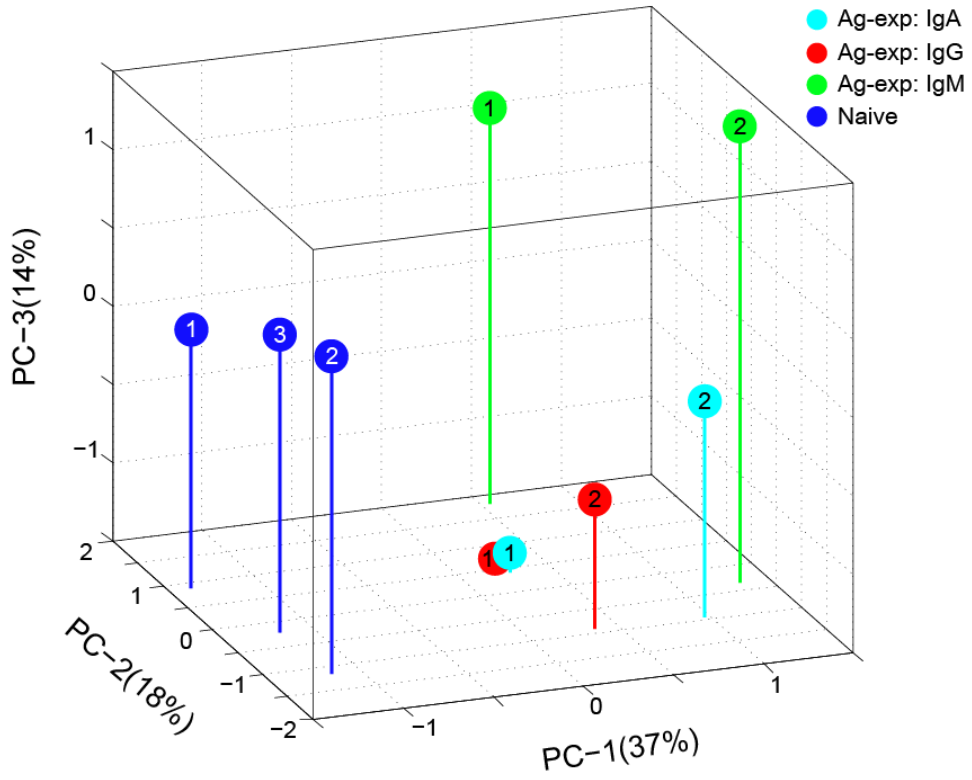

Fig. S3. Principal component analysis representation of paired VH:VL gene usage. Principal component analysis (PCA) was performed on the V-gene usage hierarchical clustering in Figure 2B; data point numbers indicate donor identification number.

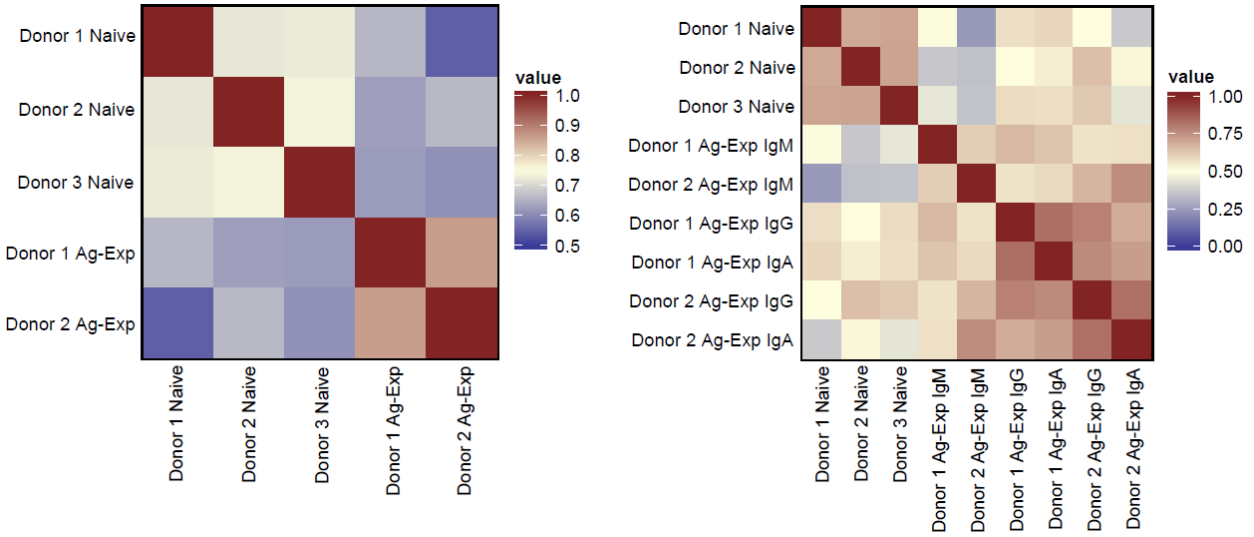

Fig. S4. Heat maps of VH:VL gene usage Pearson correlation coefficients. The Pearson hierarchical clustering analyses presented in Figure 2B were plotted as pair-wise Pearson correlation coefficient values. *Left* Naive repertoires across donors were less correlated than antigen-experienced repertoires between donors. *Right* When subsets were subdivided by isotype, the class-switched repertoires (IgG/IgA) showed the highest correlations within and across donors.

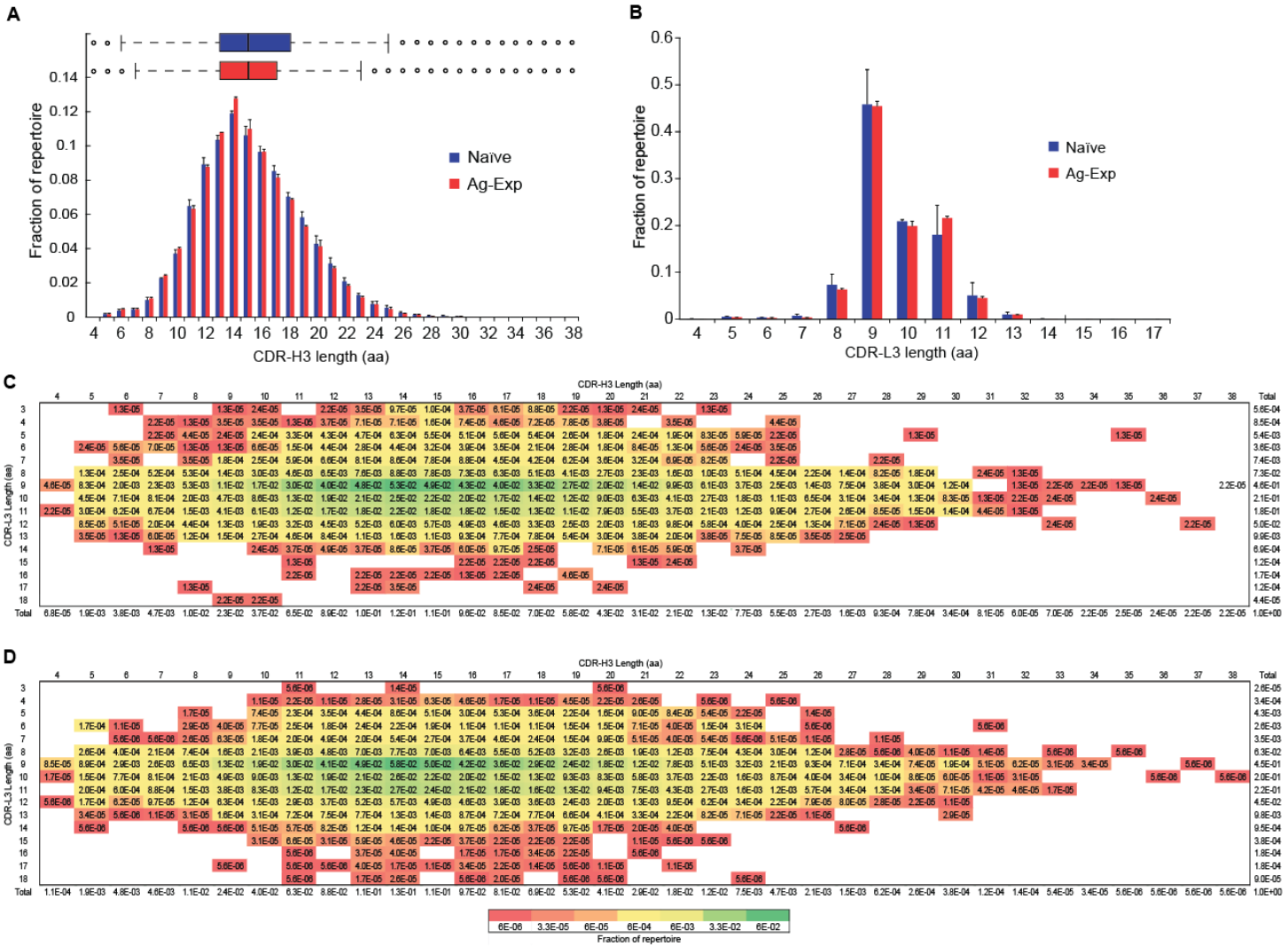

Fig. S5. **Distribution of CDR-H3 and CDR-L3 loop lengths**. *(A)* Histograms of CDR-H3 amino acid length, and *(B)* CDR-L3 amino acid length. *(C)* and *(D)* CDR-H3:CDR-L3 length heat maps of *(C)* naïve donor repertories, and *(D)* antigen-experienced donor repertoires. Data were averaged across all donors, with error bars indicating standard deviations. (A) $p < 10^{-14}$ by K-S test, which compares the equality of distributions. CDR-H3 length means: Naïve 15.21 aa, Ag-Exp 15.06 aa; medians Naïve 15, Ag-Exp 15 aa, respectively. (B) K-S test $p = 3.1\times10^{-5}$, CDR-L3 length means: Naïve 9.700, Ag-Exp 9.724; medians Naïve 9, Ag-Exp 9 aa, respectively.

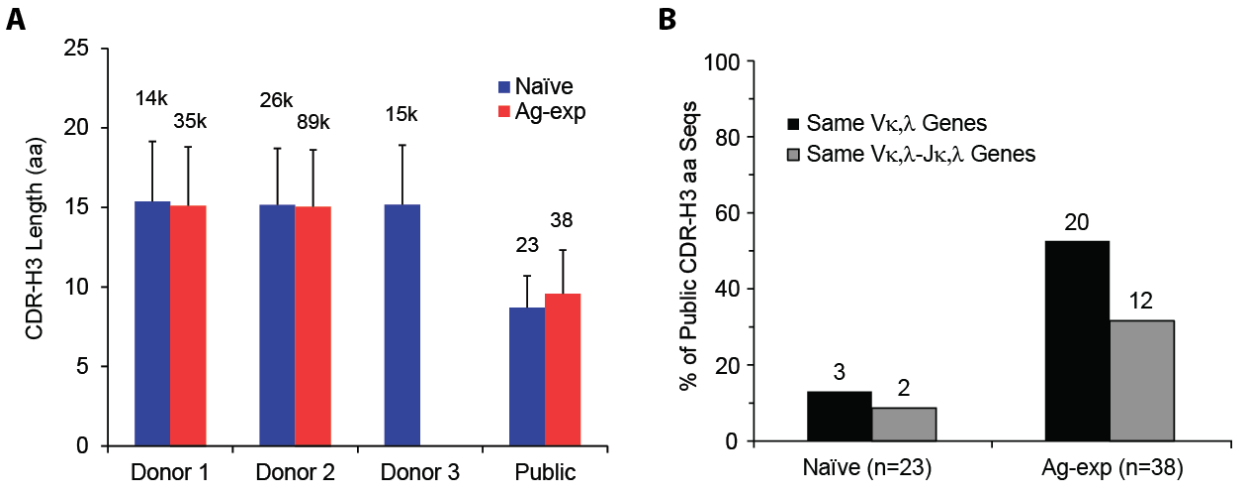

Fig. S6. Public CDR-H3 usage. *(A)* CDR-H3 length comparisons between overall repertoires and public CDR-H3 amino acid sequences. Values above each column indicate the total number of CDR-H3 in each group. *(B)* Light chain gene usage comparisons for public CDR-H3 amino acid antibodies. Public antibodies expressing the same CDR-H3 amino acid sequence displayed higher Vκ,λ and Vκ,λ-Jκ,λ gene usage convergence in antigen-experienced groups compared to naïve groups. Values above each column indicate the total number of public CDR-H3 antibodies in each group. All error bars represent standard deviation. The abbreviation *k* indicates ×1,000.

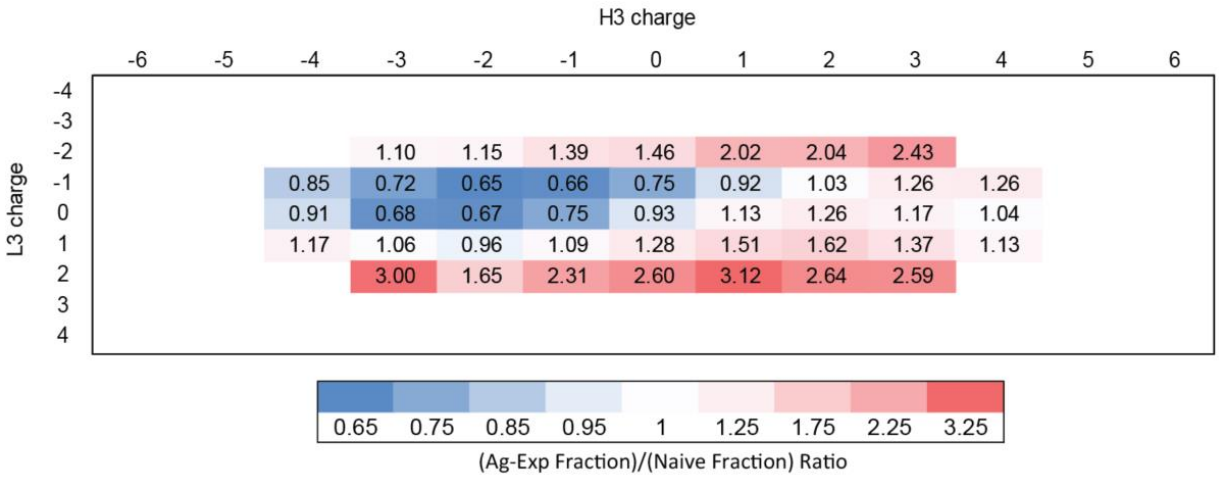

Fig. S7. CDR3 charge distribution ratio heat maps. Heat map of CDR-H3:CDR-L3 charge pair combinations across NBC and AEBC repertoires. Values represent the average ratio of antigen-experienced:naïve repertoire fractional representation for a given H3:L3 charge combination in Donors 1 and 2. Red and blue shading indicates a fractional increase or decrease in antigen-experienced compared to naïve repertoires, respectively.

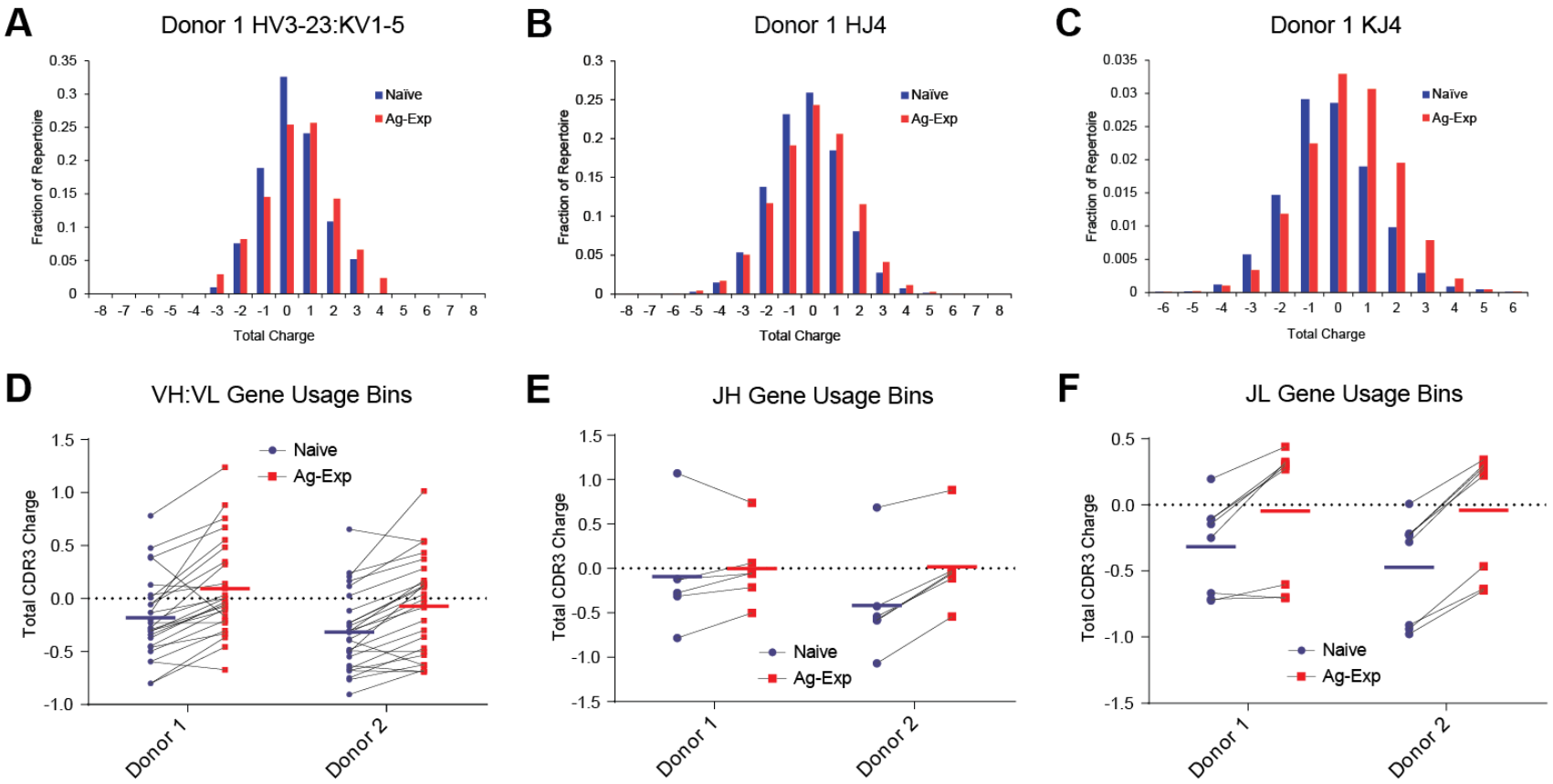

Fig. S8. CDR3 charge analysis binned by VH, VL, JH, and JL gene usage. Charge differences between naïve and antigen-experienced datasets were binned by antibody genes to understand how CDR3 charge selection occurred when controlled for gene usage. Three example gene bin histograms are shown for Donor 1 (VH:VL genes *HV3-23*:*KV1-5*, JH gene *HJ4*, and JL gene *KJ4*, in panels *(A)*, *(B)*, and *(C)*, respectively); the means of each of these distributions comprise a single point on the graphs below. *(D)*, *(E)*, and *(F)* Total CDR3 charge means were plotted for gene sets with 50 or more sequenced antibodies in all naïve and antigen-experienced Donor 1 or 2 datasets. Each point represents the mean of at least 50 antibodies that comprised that gene set. Black lines connect the same gene set across naïve and antigen-experienced repertoires, and colored lines report the mean of displayed gene sets (with equal weighting for each gene set regardless of the number of antibodies in that set). *(D)* Total CDR3 charge means of VH:VL gene usage bins (n = 26 VH:VL gene pairs with at least 50 antibodies in all datasets). *(E)* Total CDR3 charge means for JH gene usage bins (n = 5 JH genes). *(F)* Total CDR3 charge means of JL gene usage bins (n = 8 Jκ/Jλ genes).

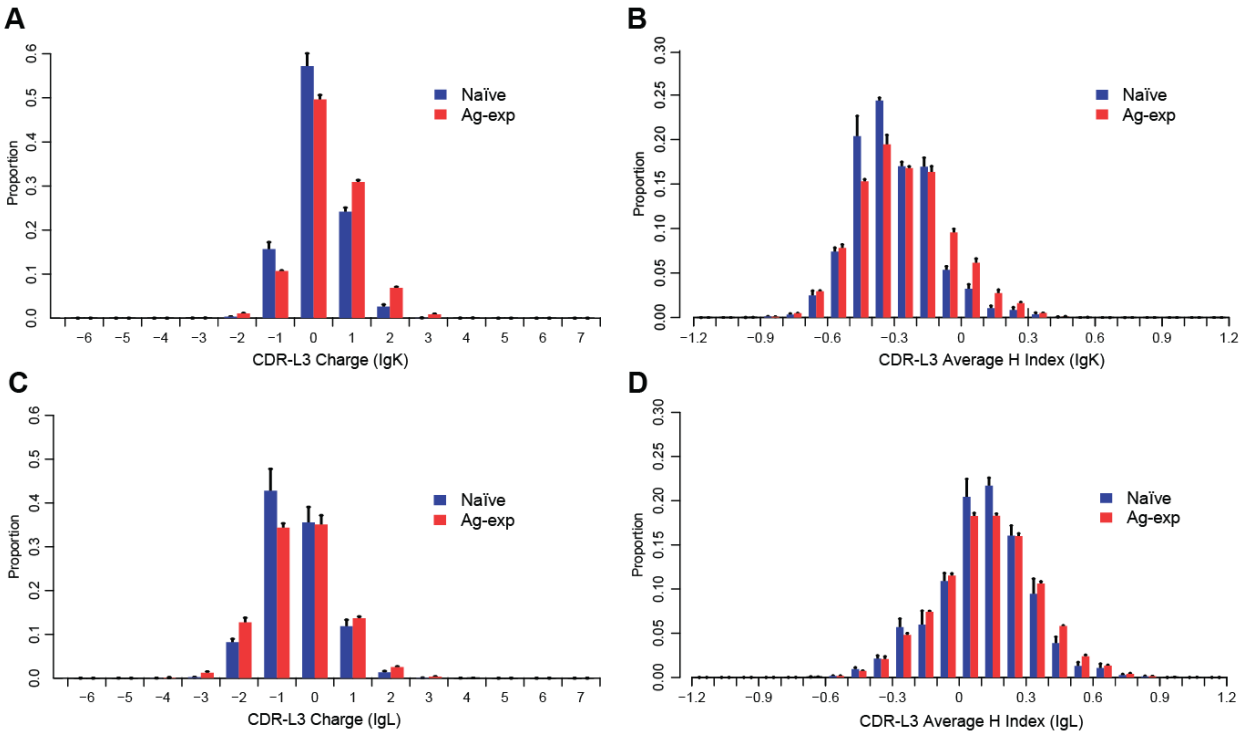

Fig. S9. Charge and hydrophobicity in kappa and lambda light chains. Igκ (*(A)* and *(B)*) and Igλ (*(C)* and *(D)*) CDR-L3 charge (*(A)* and *(C)*) and CDR-L3 average hydrophobicity (*(B)* and *(D)*). Kappa and lambda repertoires exhibited distinct CDR-L3 charge and average hydrophobicity (upper compared to lower graphs). Specifically, Igκ light chains were more likely to hold a positive CDR3 charge than Igλ (*(A)* vs. *(C)*) and Igκ exhibited lower average CDR3 hydrophobicity than Igλ (*(B)* vs. *(D)*). All naïve repertoire distributions were statistically significant from antigen-experienced repertoires by K-S test ($p < 10^{-14}$). *N* for the above repertoires is provided in Table S1; means for the above graphs are provided in Table S5.

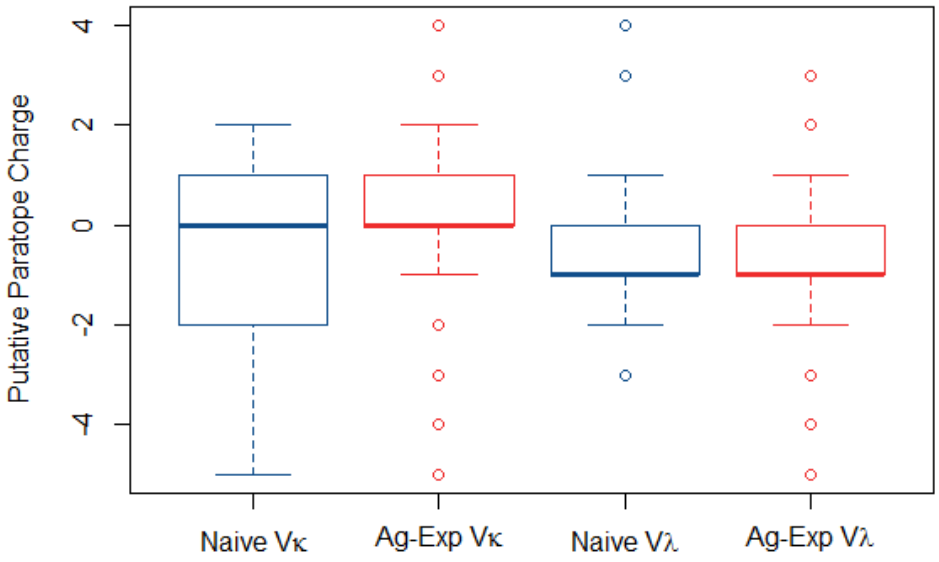

Fig. S10. Charge in kappa vs. lambda light chain CR-paratopes. Igκ and Igλ CR-paratope charge in naïve and antigen-experienced antibody repertoires is shown. All naïve kappa versus lambda distributions were statistically significant by the K-S test (p = 1.8 x 10^{-9} for naïve and $p < 10^{-15}$ for antigen-experienced distributions).

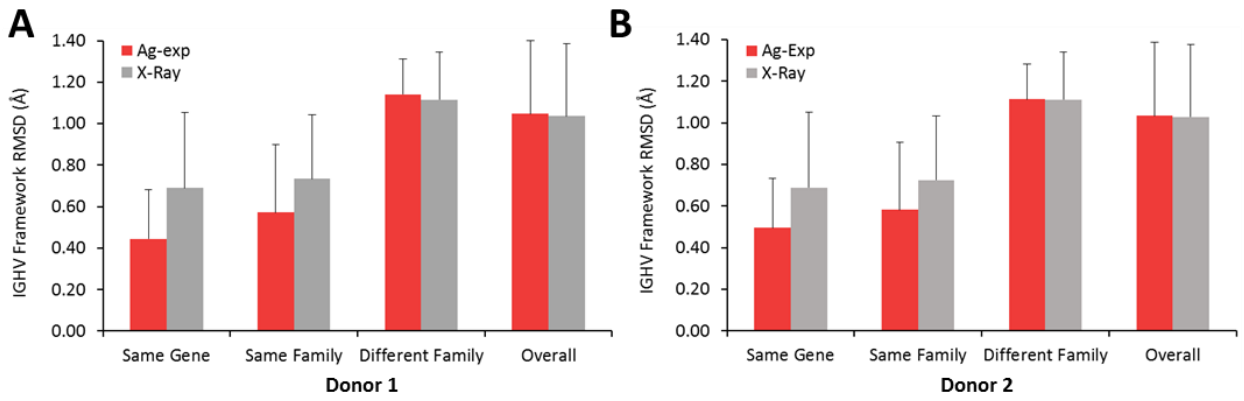

Fig. S11. Average root mean square deviation (RMSD) of framework backbone atoms in antigen-experienced repertoires compared to unique human antibody solved crystal structures. RMSDs of all antibodies in a repertoire to each other were calculated for V_H framework 1-3 backbone atoms and binned by V-gene usage. *(A)* Donor 1 and *(B)* Donor 2 RMSD of antibodies sharing the same V-gene segment; same V-gene family; different V-gene family; and overall repertoire comparison for antigen-experienced (red) and crystal structure (gray) repertoires. $p < 10^{-15}$ for all antigen-experienced versus solved crystal structure RMSD distributions. *N* for antigen-experienced repertoires is provided in Table S1; *N*=141 for PDB crystal structures.